Examining the Effect of Biomarkers in Terms of Compartmentalization and a Continuous Variable

Irene B. Helenowski\textsuperscript{1,2}, Hakan Demirtas\textsuperscript{2}, Edward F. Vonesh\textsuperscript{1,3}, Ryan J. Deaton\textsuperscript{4}, Borko D. Jovanovic\textsuperscript{1,5}, Alfred W. Rademaker\textsuperscript{1}, Sally A. Freels\textsuperscript{2}, Vijayalakshmi Ananthanarayanan\textsuperscript{4}, Peter H. Gann\textsuperscript{4}

Department of Preventive Medicine, Feinberg School of Medicine, Robert H. Lurie Comprehensive Cancer Center, Northwestern University, Chicago, IL\textsuperscript{1}

Division of Epidemiology & Biostatistics, School of Public Health, University of Illinois – Chicago, Chicago, IL\textsuperscript{2}

Baxter Healthcare Corporation, Round Lake, IL\textsuperscript{3}

Department of Pathology, University of Illinois – Chicago, Chicago, IL\textsuperscript{4}

General Clinical Research Center, Feinberg School of Medicine, Northwestern University, Chicago, IL\textsuperscript{5}
PURPOSE

- Here, we address models used to examine the effect of spatial distance and pathological compartmentalization accounting for the issue of paucity in the data.

- Due to this paucity, we have had trouble in convergence of some models.

- We therefore attempt to apply these models to imputed data created using either:
  - A Bootstrap approach
  - A Bayesian-like approach

- Our data consists of *MCM2* index values coming from the prostate gland.
The data comes from ten subjects, ages 47-68, having prostate biopsies between 2002 and 2003.

Figure A shows a diagram of how the prostate gland zones from which the data were collected were designated.

The numbers in parentheses in Figure A correspond to the numbering of the biopsy needle cores assigned to take a sample from that particular zone. Each prostate zone can include a normal compartment, an abnormal compartment, or both.

The numbers in brackets give the coordinates of each section used in computation of the Euclidean distance. For coordinates \([x, y]\), for example, Euclidean distance \(=\sqrt{x^2 + y^2}\). 


1. BR = Basal Right section [4,2]
2. M1-R = Middle 1 Right section [3,2]
3. M2-R = Middle 2 Right section [2,2]
4. AR = Apical Right section [1,2]
5. TZ Zone [not in analysis]
6. BL = Basal Left section [4,1]
7. M1-L = Middle 1 Left section [3,1]
8. M2-L = Middle 2 Left section [2,1]
9. AL = Apical Left section [1,1]
10. TZ Zone [not in analysis]
We define the **MCM2 (minichromosome maintenance protein 2) index** by:

- Calculating the percentage of cells positively either strongly or weakly staining for that biomarker from the total number of basal cells in each prostate gland zone or:
  - Calculating the percentage of cells positively strongly staining for that biomarker from the total number of basal and luminal cells in each prostate gland zone.
Our statistical analyses involved the following models that are used to:

- Obtain Least Square Differences between different needle cores, each core corresponding to one prostate gland zone, and pathological compartments.

\[
y_{ijk} = \mu + \text{Subject}_i \text{+ Needle-Core}_j \text{+ Pathological_compartment}_{k(j)} + \text{error}_{ijk} \quad (1)
\]

for \( i = 1, \ldots, 10 \) subjects
\( j = 1, \ldots, 8 \) needle cores

and \( \text{Pathological_compartment}_k = \begin{cases} 1, & \text{compartment is abnormal}, (k=1) \\ 0, & \text{compartment is normal}, (k=2) \end{cases} \)
MODELS USED

- And predict differences based on distance and compartmentalization.

\[
\text{LSDifference}_{qr} = \mu + \text{distance}_{qr} + \text{Pathological}_\text{compartment}_{qr} \\
+ (\text{distance} \times \text{Pathological}_\text{compartment})_{qr} + \text{error}_{qr}
\]  

(2)

for \( q = 1, 2, 3, 4, 6, 7, 8 \) and \( r = 2, 3, 4, 6, 7, 8, 9 \)

where \( q \) and \( r \) indexes the first and second number of the needle cores, respectively, associated with a particular pair of MCM2 samples from which the least-square differences are calculated.

- In Model (2), \( \text{distance}_{qr} \) is the Euclidean distance between the prostate gland zones associated with needle cores \( q \) and \( r \) and:

\[
\text{Pathological}_\text{compartment}_{qr} = \begin{cases} 
-1, \text{Normal} - \text{Abnormal \_difference} \\
0, \text{Same \_Compartment \_differences} \\
1, \text{Abnormal} - \text{Normal \_differences}
\end{cases}
\]
The responses in these models are the rank orders of the MCM2 index.

These models were applied to the data set assuming compound symmetry.
RESULTS
(COMPLETE CASE ANALYSES)

- Figure B shows the least square differences within distance between prostate gland zones:
  - Associated with cores from the same compartments
  - And cores of normal compartments subtracted from cores of abnormal compartments.
RESULTS
(COMPLETE CASE ANALYSES)

PLOTS OF LEAST-SQUARE DIFFERENCES IN MCM2 INDEX VALUES BETWEEN PROSTATE GLAND ZONES AVERAGED OVER EUCLIDEAN DISTANCE BETWEEN THE ZONES

Based on first definition

Based on second definition
RESULTS
(COMplete CASE ANALYSES)

- Given the first definition for the MCM2 index, the differences based on subtractions of abnormal compartment data from normal compartment data are lower than differences between data from the same compartments.

- Given the second definition for the MCM2 index, the differences based on subtractions of abnormal compartment data from normal compartment data are higher than differences between data from the same compartments.
RESULTS  
(COMplete-cASE  
ANALySES)

- Table 1 (below) shows the *p*-values coming from regression analyses when the least-square differences for the MCM2 index given either definition are applied to Model (2).
- While the effect of compartmentalization on least square differences is significantly different at 5% (*p* < 0.01) given either definition of the MCM2 index, distance and the interaction between distance and compartmentalization do not significantly affect least square differences.

<table>
<thead>
<tr>
<th><em>p</em>-value coming from Model (2) applied to LS Differences for the MCM2 index based on:</th>
<th>Euclidean Distance</th>
<th>Pathological Compartmentalization</th>
<th>Euclidean Distance * Pathological Compartmentalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Definition</td>
<td>0.46</td>
<td>&lt; 0.01</td>
<td>0.67</td>
</tr>
<tr>
<td>Second Definition</td>
<td>0.86</td>
<td>&lt; 0.01</td>
<td>0.83</td>
</tr>
</tbody>
</table>
Frequency of missing vs. non-missing values.

- “Ideal” completely balanced data set would have
  \[ \text{2 (pathological compartmentalization types)} \times \]
  \[ \text{8 (biopsy needle cores)} \times \]
  \[ \text{10 (subjects)} = 160 \text{ observations} \]

- Thus, each pathological compartmentalization type would have 80 observations.
In reality:

- We have 74 observations (i.e., 6 missing values) for the normal compartment type and only 10 observations (i.e., 70 missing values) for the abnormal compartment type given the first definition of the MCM2 index.

- We have 74 observations (i.e., 6 missing values) for the normal compartment type and only 50 observations (i.e., 30 missing values) for the abnormal compartment type given the second definition of the MCM2 index.
Multiple imputation involving bootstrapping was based on bootstrapping predictive values obtained from Model (1) applied to the entire data set under the assumption of compound symmetry.

Boostrapping with either definition for the MCM2 index was conducted by:

- Either ignoring pathological compartmentalization in which predicted values were sampled to fill in for missing values regardless of their pathological type.
- Or considering pathological compartmentalization in which predicted values to be filled in for the missing data were sampled so that a sampled predicted value came from the same pathological compartment type as the one associated with that missing observation.
HANDLING PAUCITY OF DATA (BAYESIAN PREDICTIVE DISTRIBUTION)

- Data was also augmented using an approach based on the Bayesian predictive distribution also by either ignoring or considering pathological compartmentalization.
- For the first definition:
  - With ignoring pathological compartmentalization, values to fill in for missing observations were drawn from prior distributions:
    - $N(34, 8)$, $N(34, 64)$, and $N(34, 1)$, where the value 34 is the mean of the observations ignoring pathological compartmentalization.
  - With considering pathological compartmentalization, values to fill in for missing observations were drawn from prior distributions:
    - $N(30, 10)$, $N(30, 100)$, and $N(30, 1)$ for the abnormal pathological compartments, where the value 30 is the mean of the observations from the abnormal compartments.
    - $N(35, 7)$, $N(35, 49)$, and $N(35, 1)$ for the normal pathological compartments, where the value 35 is the mean of the observations from the normal compartments.
For the second definition:

- With ignoring pathological compartmentalization, values to fill in for missing observations were drawn from prior distributions:
  - \( \mathcal{N}(19, 7) \), \( \mathcal{N}(19, 49) \), and \( \mathcal{N}(19, 1) \), where the value 19 is the mean of the observations ignoring pathological compartmentalization.

- With considering pathological compartmentalization, values to fill in for missing observations were drawn from prior distributions:
  - \( \mathcal{N}(23, 7) \), \( \mathcal{N}(23, 49) \), and \( \mathcal{N}(23, 1) \) for the abnormal pathological compartments, where the value 23 is the mean of the observations from the abnormal compartments.
  - \( \mathcal{N}(16, 4) \), \( \mathcal{N}(16,16) \), and \( \mathcal{N}(16, 1) \) for the normal pathological compartments, where the value 16 is the mean of the observations from the normal compartments.
Imputed data were created with both approaches, each approach either ignoring or considering pathological compartmentalization in sets of $M = 1000$, 10, or 5 imputations, where missingness is assumed to depend on observed data only.
RESULTS

- Table 2 shows the $p$-value ranges coming from Model (2) applied to three sets separately, with each set involving 1000 bootstrap simulations either ignoring or considering pathological simulation.

- Each simulation involved filling in missing values with random draws from the predicted values obtained by applying Model (1) to all the non-missing data based on either the first or second definition.

- Imputed data were again applied to Model (1) to obtain the least-square differences were then averaged across the imputation sets within each core and pathological compartment and these averages were then applied to Model (2).
RESULTS

Table 2: *p*-values coming from regression analyses when the least-square differences for the MCM2 index given the first and second definitions are applied to Model (2) in the analyses of bootstrap imputed data.

<table>
<thead>
<tr>
<th>p-value based on method involving:</th>
<th>Euclidean Distance</th>
<th>Pathological Compartmentalization</th>
<th>Euclidean Distance * Pathological Compartmentalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple Imputation Ignoring Pathological Compartmentalization Given:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First Definition</td>
<td>0.70 – 0.82</td>
<td>0.01 – 0.03</td>
<td>0.70 – 0.90</td>
</tr>
<tr>
<td>Second Definition</td>
<td>0.19 – 0.21</td>
<td>&lt; 0.01</td>
<td>0.24 – 0.26</td>
</tr>
<tr>
<td>Multiple Imputation Considering Pathological Compartmentalization Given:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First Definition</td>
<td>0.77 – 0.79</td>
<td>0.01 – 0.05</td>
<td>0.73 – 0.94</td>
</tr>
<tr>
<td>Second Definition</td>
<td>0.34 – 0.35</td>
<td>&lt; 0.01</td>
<td>0.53 – 0.54</td>
</tr>
</tbody>
</table>
RESULTS

- Table 3 shows the $p$-value ranges coming from Model (2) applied to three sets separately, with each set involving 1000 simulations based on a Bayesian predictive distribution ignoring pathological simulation.

- Each simulation involved filling in missing values with random draws from the prior distribution given below based on first and second definitions of the MCM2 index.

- Imputed data were again applied to Model (1) to obtain the least-square differences were then averaged across the imputation sets within each core and pathological compartment and these averages were then applied to Model (2).
## RESULTS

Table 3: $p$-values coming from regression analyses when the least-square differences for the MCM2 index given the first and second definitions are applied to Model (2) in the analyses of data imputed via a Bayesian predictive distribution ignoring compartmentalization.

<table>
<thead>
<tr>
<th>Bayesian predictive Distribution assumed for both compartments</th>
<th>Euclidean Distance</th>
<th>Pathological Compartmentalization</th>
<th>Euclidean Distance * Pathological Compartmentalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple Imputation Ignoring Pathological Compartmentalization Given First Definition</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N[34, 8]</td>
<td>0.72 – 0.76</td>
<td>0.03 – 0.05</td>
<td>0.86 – 0.91</td>
</tr>
<tr>
<td>N[34, 64]</td>
<td>0.73 – 0.90</td>
<td>&lt; 0.01 – 0.02</td>
<td>0.75 – 0.83</td>
</tr>
<tr>
<td>N[34, 1]</td>
<td>0.54 – 0.57</td>
<td>0.67 – 0.69</td>
<td>0.97 – 0.99</td>
</tr>
<tr>
<td>Multiple Imputation Ignoring Pathological Compartmentalization Given Second Definition</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N[19, 7]</td>
<td>0.17 – 0.19</td>
<td>&lt; 0.01</td>
<td>0.24 – 0.25</td>
</tr>
<tr>
<td>N[19, 49]</td>
<td>0.20 – 0.21</td>
<td>&lt; 0.01</td>
<td>0.25 – 0.27</td>
</tr>
<tr>
<td>N[19, 1]</td>
<td>0.19 – 0.21</td>
<td>&lt; 0.01</td>
<td>0.16 – 0.18</td>
</tr>
</tbody>
</table>
RESULTS

- Table 4 shows the $p$-value ranges coming from Model (2) applied to three sets separately, with each set involving 1000 simulations based on a Bayesian predictive distribution considering pathological simulation.

- Each simulation involved filling in missing values with random draws from the prior distributions given below corresponding to the normal and abnormal pathological compartments, respectively, based on first and second definitions of the MCM2 index.

- Imputed data were again applied to Model (1) to obtain the least-square differences were then averaged across the imputation sets within each core and pathological compartment and these averages were then applied to Model (2).
Table 4: *p*-values coming from regression analyses when the least-square differences for the MCM2 index given the first and second definitions are applied to Model (2) in the analyses of data imputed via a Bayesian predictive distribution considering compartmentalization.

<table>
<thead>
<tr>
<th>Bayesian predictive distribution for Normal Compartment</th>
<th>Bayesian predictive distribution for Abnormal Compartment</th>
<th>Euclidean Distance</th>
<th>Pathological Compartmentalization</th>
<th>Euclidean Distance * Pathological Compartmentalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple Imputation Considering Pathological Compartmentalization Given First Definition</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N[35, 7]</td>
<td>N[30, 10]</td>
<td>0.67 – 0.75</td>
<td>&lt; 0.01</td>
<td>0.97 – 0.98</td>
</tr>
<tr>
<td>N[35, 49]</td>
<td>N[30, 100]</td>
<td>0.75 – 0.84</td>
<td>&lt; 0.01</td>
<td>0.79 – 0.87</td>
</tr>
<tr>
<td>N[35, 1]</td>
<td>N[30, 1]</td>
<td>0.36 – 0.37</td>
<td>&lt; 0.01</td>
<td>0.89 – 0.90</td>
</tr>
<tr>
<td>Multiple Imputation Considering Pathological Compartmentalization Given Second Definition</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N[16, 4]</td>
<td>N[23, 7]</td>
<td>0.32 – 0.34</td>
<td>&lt; 0.01</td>
<td>0.56 – 0.59</td>
</tr>
<tr>
<td>N[16, 16]</td>
<td>N[23, 49]</td>
<td>0.24 – 0.27</td>
<td>&lt; 0.01</td>
<td>0.78 – 0.81</td>
</tr>
<tr>
<td>N[16, 1]</td>
<td>N[23, 1]</td>
<td>0.42 – 0.43</td>
<td>&lt; 0.01</td>
<td>0.76 – 0.79</td>
</tr>
</tbody>
</table>
RESULTS

- For $M = 5$ and $M = 10$ imputations, results were exactly the same as for $M = 1000$ in the bootstrap approach and similar in the Bayesian approach; the graph below shows the average least-square differences over Euclidean distance for each of $M = 5$ imputations.

First Definition

Second definition

![Graphs showing least-square mean differences over Euclidean distance for different definitions and approaches with and without considering pathological compartmentalization.]
CONCLUSION

- Given the results, we can see that the similarity of our imputation method and the results from the complete-case analyses depends on the type of method.

- The consistency also appears to depend on the number of missing values in the original data.

- More investigation is needed to assess how the imputation method affect variance estimates, essential for inference of parameter estimates and future biomarker reproducibility studies.
References


Recommended Literature


Acknowledgements

This work was supported by the Prostate SPORE (NIH/NCI P50 CA 90386) and Breast SPORE (P50 CA 89018) grants as well as the NIH/NCI R01 CA 90759 grant through the Robert H. Lurie Comprehensive Cancer Center of Northwestern University.